(12) UK Patent Application (19) GB (11) 2 346 527 (13) A

(43) Date of A Publication 09.08.2000

(21) Application No 0009707.1

(22) Date of Filing 20.07.1998

Date Lodged 19.04.2000

(30) Priority Data
(31) 08900810 (32) 25.07.1997 (33) US

(62) Divided from Application No 9815620.1 under Section 15(4) of the Patents Act 1977

(51) INT CL[7]
G06T 1/40 15/70

(52) UK CL (Edition R)
H4T TABX

(56) Documents Cited
EP 0696018 A2

(58) Field of Search
UK CL (Edition R) H4T TABX
INT CL[7] G06T 1/40 15/70

(71) Applicant(s)
Motorola Inc
(Incorporated in USA - Delaware)
1303 East Algonquin Road, Schaumburg,
Illinois 60196, United States of America

(72) Inventor(s)
Noel Massey
Orhan Karaali
Otto Schnurr

(74) Agent and/or Address for Service
Motorola Limited
European Intellectual Property Department, Midpoint,
Alencon Link, BASINGSTOKE, Hampshire, RG21 7PL,
United Kingdom

(54) Abstract Title
Virtual actor with set of speaker profiles

(57) A set of speaker profiles are provided, including at least visual profile data and voice profile data. These cause a virtual actor to be displayed 514 with a visual appearance determined by the visual profile data and a voice determined by the voice profile data.

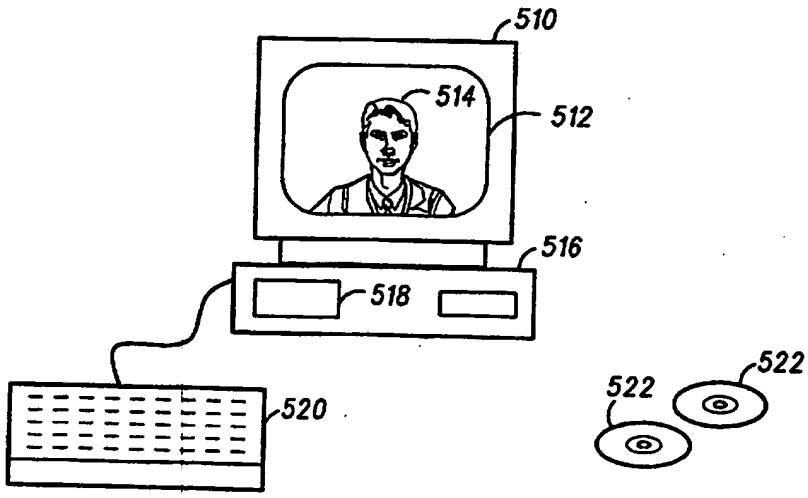

FIG.5

GB 2 346 527 A

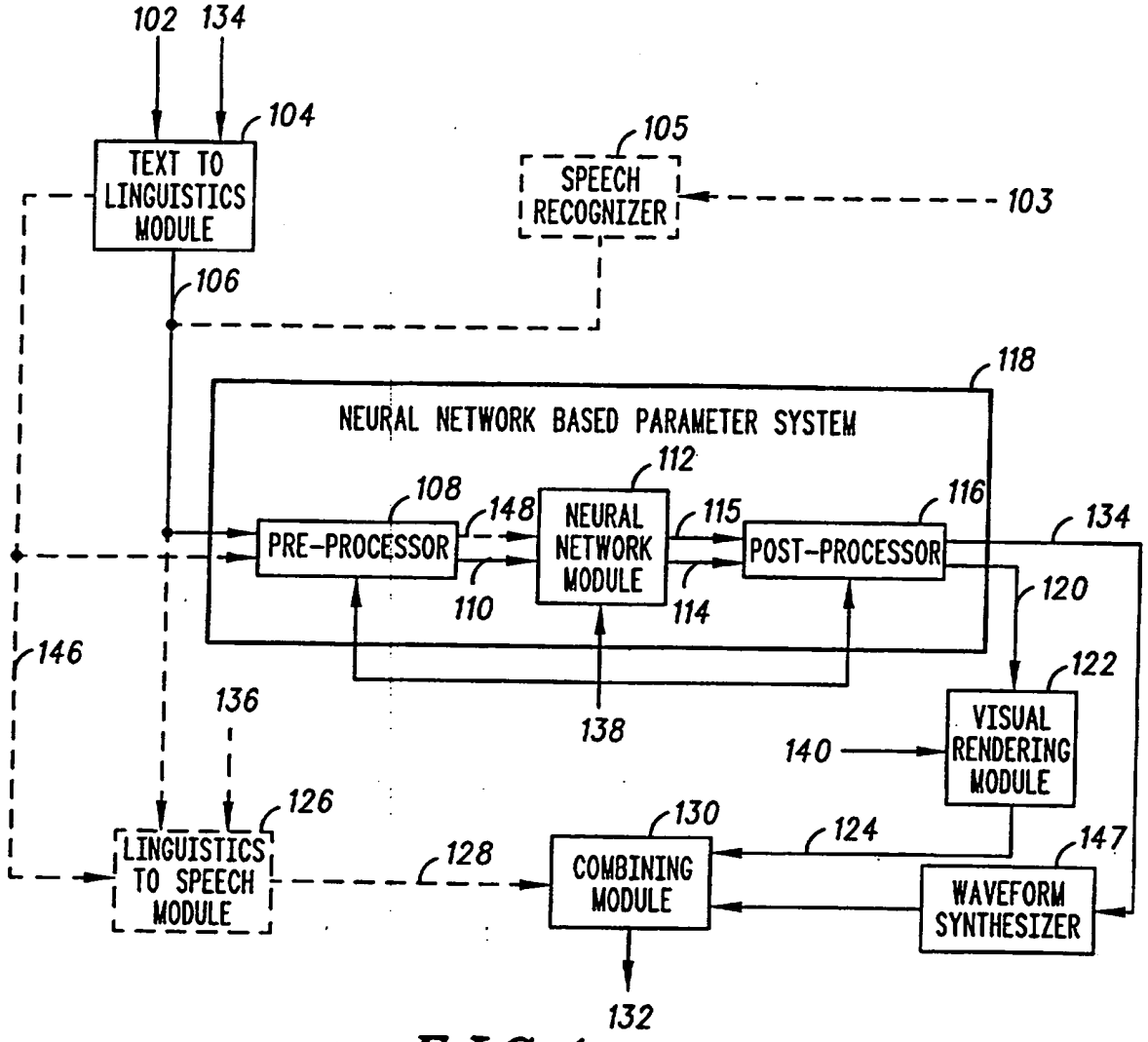
102
134
104
TEXT TO LINGUISTICS MODULE
105
SPEECH RECOGNIZER
103
106
118
NEURAL NETWORK BASED PARAMETER SYSTEM
112
108
148
NEURAL NETWORK MODULE
115
116
PRE-PROCESSOR
POST-PROCESSOR
134
110
114
120
146
122
136
138
VISUAL RENDERING MODULE
140
126
130
124
147
LINGUISTICS TO SPEECH MODULE
128
COMBINING MODULE
WAVEFORM SYNTHESIZER
132

FIG.1

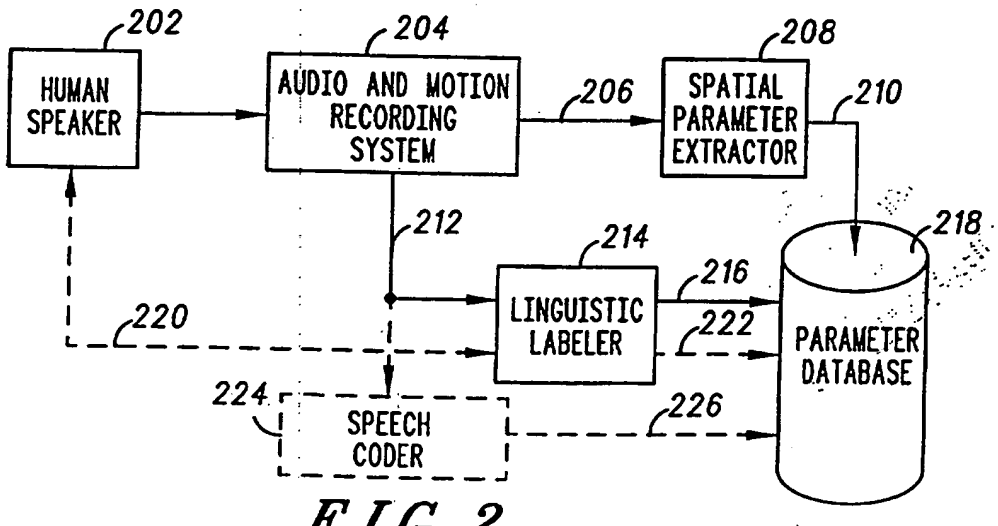
202
204
208
HUMAN SPEAKER
AUDIO AND MOTION RECORDING SYSTEM
206
SPATIAL PARAMETER EXTRACTOR
210
212
214
218
216
LINGUISTIC LABELER
PARAMETER DATABASE
220
222
224
226
SPEECH CODER

FIG.2

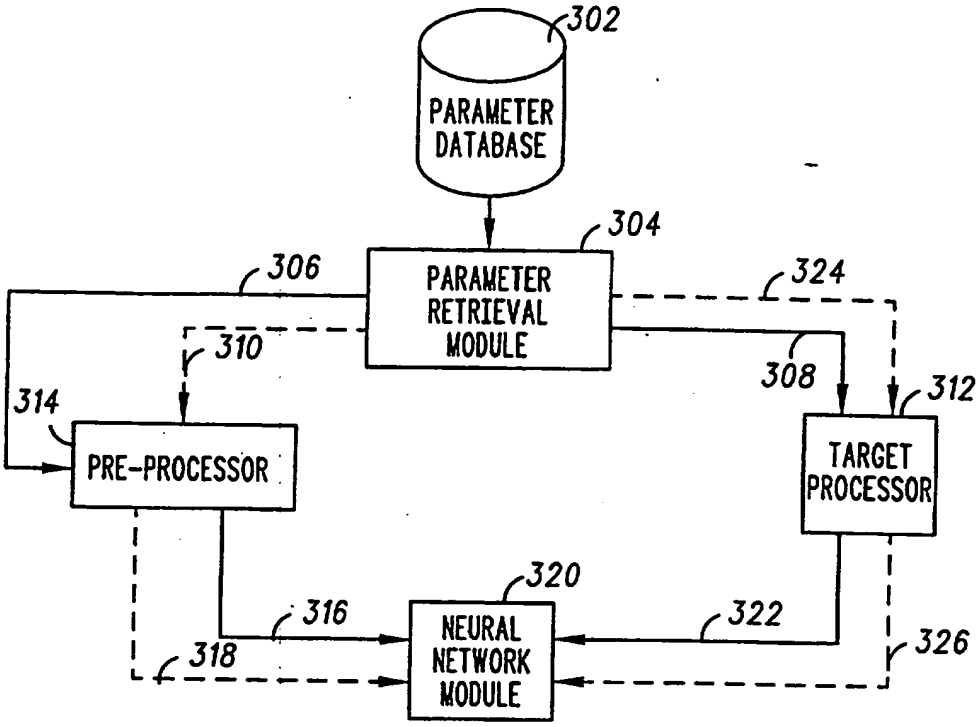
302
PARAMETER DATABASE
304
306
PARAMETER RETRIEVAL MODULE
324
310
308
312
314
PRE-PROCESSOR
TARGET PROCESSOR
320
316
322
318
326
NEURAL NETWORK MODULE

FIG.3

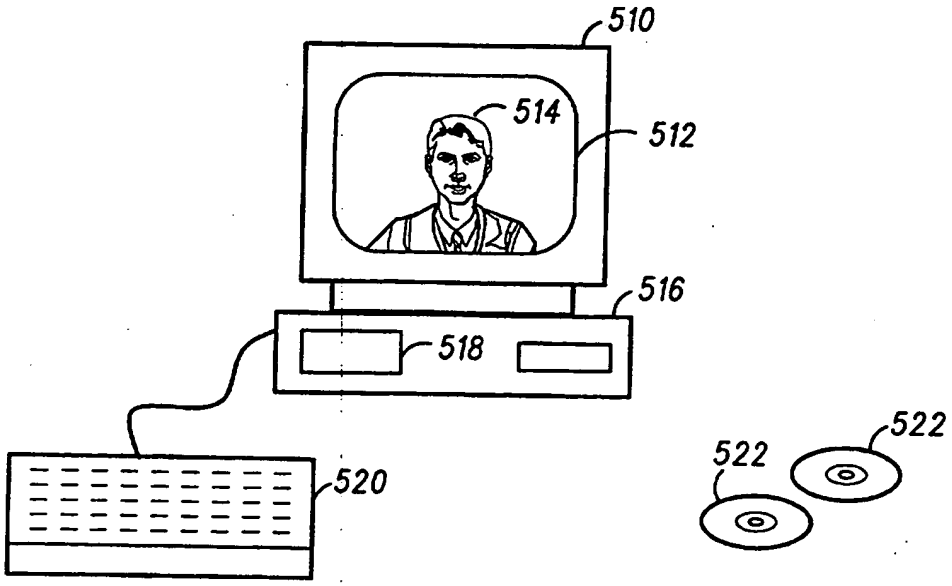
510
514
512
516
518
520
522
522

FIG.5

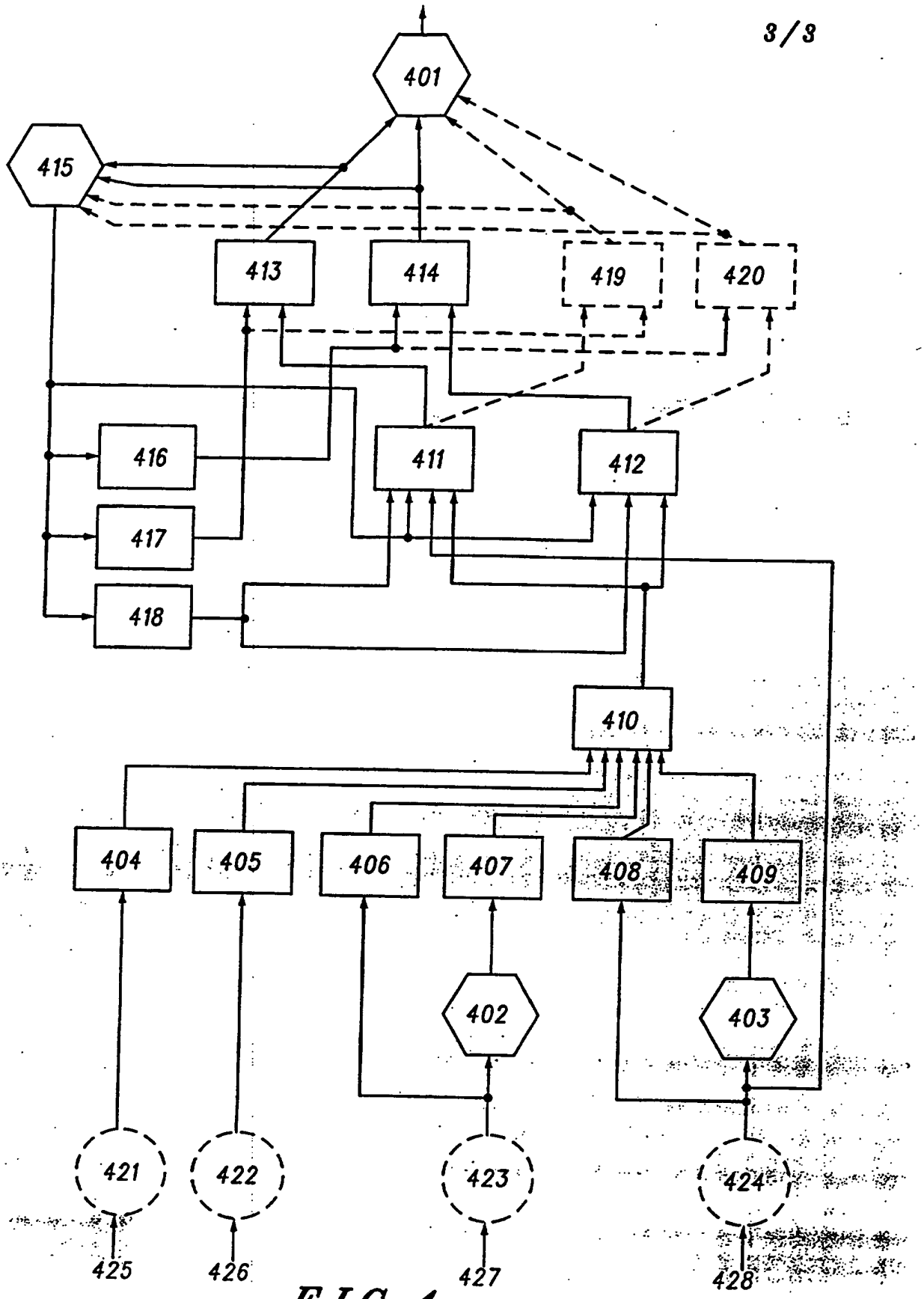
401
415
413
414
419
420
416
411
412
417
418
410
404
405
406
407
408
409
402
403
421
422
423
424
425
426
427
428

FIG. 4

2346527

VIRTUAL ACTOR WITH SET OF SPEAKER PROFILES

Field of the Invention

The present invention relates to generating model parameters for animating virtual actors.

Background of the Invention

Computer programs exist which include providing a feature that draws an animated figure which appears to be speaking text from sources such as e-mail, word processing documents and internet pages. Typically, this animated object is called a "Talking Head", though the head is seldom very realistic in its appearance or movements. By using "target frames", which are the frames in the animation that correspond to the middle of a phoneme being synthesized, fairly realistic animation can be produced but the transitions between phonemes are typically very jumpy and unnatural looking. Most of the "Talking Heads" are therefore limited in their usefulness due to their lack of realism.

There are three standard approaches to creating these "Talking Heads". The simplest approach is to store a limited number (typically between five and about one hundred) of pictures that correspond to various lip positions. These images are recalled and displayed at the appropriate time so that the "Talking Head" appears to be moving.

The second method is to store "key frames" which describe the position of the lips and other objects for target frames. Typically, these frames correspond to the middle of each of the

phonemes. The frames between the key frames are derived from the key frames by some form of interpolation. This method creates smoother animation but the interpolation is often too simplistic to create realistic motions. For example, when a person says "pop" the lips do not follow a simple linear trajectory form of closed to open to closed. The key frame method would most likely just average these locations, resulting in a "Talking Head" that still lacks the ability to realistically imitate human motions.

Moreover, for the methods that store key frames, there is a trade-off between smoothness of the animation and the memory or disk space that is required to create the "Talking Head". The storage requirement increases with the number of pictures that are stored which means that smooth animation requires substantial storage space. Systems with restricted memory and disk space are forced to use a smaller number of frames which makes the animation very jumpy and unnatural.

The third option is to write rules to describe the lip locations. These rules can produce natural motion if enough data such as the size and weight of the lips, placement and strengths of the face muscles and jaw positions can be obtained. The problem is that the rules can be very complicated and can take a long time to write. The rule based system can also be difficult to modify in order to model a new object.

At least one disadvantage with all three approaches is the cost of achieving fidelity. The frame rate at which the animation is rendered directly impacts fidelity. In order to support fast frame rates, the system must either incur a storage cost for

utilizing an extensive database or incur a development cost for manually creating a set of rules complex enough to accurately predict the control parameters of the "Talking Head".

Another disadvantage of the traditional methods is that it is often difficult to capture the personal idiosyncrasies of the specific person that is to be modeled. If it is desirable to have a "Talking Head" that imitates a person, then it is desirable to have it speak, nod and move like that person.

Hence, there is a need for a method and apparatus for improving the fidelity of autonomously rendered models without increasing the cost of animating the model.

Brief Description of the Drawings

A preferred embodiment of the invention is now described, by way of example only, with reference to the accompanying drawings in which:

FIG. 1 is a schematic representation of a neural network based parameter system for generating model parameters in accordance with the present invention;

FIG. 2 is a schematic representation of a device for generating a parameter database in accordance with the present invention;

FIG. 3 is a schematic representation of a device for training a neural network in accordance with the present invention;

FIG. 4 is a schematic representation of an embodiment of a neural network in accordance with the present invention;

FIG. 5 is a general purpose computer implementing the present invention.

Detailed Description of the Preferred Embodiment

The present invention provides a neural network based virtual actor (Talking Head) in parallel with a text-to-speech system which is used to control virtual actors (e.g., three-dimensional graphical objects). Hence, the present invention provides a method and apparatus for efficiently generating spatial parameters from linguistic information using a neural network, thus enabling the autonomous rendering of animated models with high fidelity and low cost.

It is desired to improve the quality of computer software through increased comprehension of speech synthesizers through the addition of a virtual actor which is a model of a talking human. However, the primary problem with the standard approach is that the traditional talking head does not move like a real person. A major contribution to the unnaturalness is due to the difficulty in modeling co-articulatory effects. Most models neglect both the effects of previous or future jaw and mouth positions and the complicated interactions which leads to unrealistic models. This means that there is insufficient detail to allow a deaf person to lip-read from the talking head and it also means that there is only slight improvements in intelligibility for normal hearing people.

In contrast to the traditional methods, the present invention is ideal for the task of creating a virtual actor. One advantage

of using a neural network instead of traditional methods is that the process is highly automated. Once the training points are generated for the desired person or object, the training of a neural network involves very little human interaction. In this process, the neural network is capable of capturing the idiosyncrasies of the subject, including facial gestures, head movements, shoulder shrugs, waving, and other gestures of the hands and face. In addition, since the neural network is trained on data from a real speaker, the motions will be realistic and natural. The neural network will also be able to learn the appropriate co-articulatory effects which will further enhance the naturalness of the virtual actor.

Referring to the drawings, FIG. 1 is a schematic representation of a neural network based parameter system for generating model parameters in accordance with the present invention. In a preferred embodiment, the neural network based parameter system 118 is used to generate a virtual actor (visually rendered model with speech 132) whose movements are correlated with synthetic speech (neural network synthesized speech 144). Text 102 is used to drive the virtual actor and the output is an animated object that is synchronized with synthetic speech (visually rendered model with speech 132). Text 102 is first converted to a linguistic representation of speech 106 by a text-to-linguistics module 104. The linguistic representation of speech 106 describes speech segments which are preferably phoneme segments. The text, therefore, consists of a series of segment descriptions which are composed of the following: a phoneme

identifier associated with a phoneme in current and adjacent segment descriptions; articulatory features associated with the phoneme in current and adjacent segment descriptions; locations of syllables, words and other syntactic and intonational boundaries; duration of time between phonemes, syllables, words and other syntactic and intonational boundaries; syllable strength information; descriptive information of a word type; and prosodic information. The prosodic information further includes at least one of the following: locations of word endings and a degree of disjuncture between words, locations of pitch accents and a form of those accents, locations of boundaries marked in a pitch contour and a form of those boundaries, a time separating marked prosodic events from other marked prosodic events, and a number of prosodic events of some type in a time period separating a prosodic event of another type and a frame for which coder parameters are generated.

The linguistic representation of speech 106 is converted to neural network linguistic parameters 110 by a pre-processor 108. It should be noted that the neural network linguistic parameters differ from acoustic parameters (such as linear prediction coefficients, line spectral frequencies, spatial energy parameters, pitch, etc.). The pre-processor 108 generates the neural network linguistic parameters 110 which are numbers that are suitable for use by a neural network. A neural network module 112 is used to convert the neural network linguistic parameters 110 into raw speech parameters 115 and raw spatial parameters 114. Since the output of the neural network module 112 only generates

values in the range of zero and one, a post-processor 116 is used to convert the raw spatial parameters 114 and raw speech parameters 115 into model parameters 120 and a parametric representation of speech 134, respectively. The post-processor 116 scales the output of the neural network module 112 and converts the output into parameters that are suitable for driving a visual rendering module 122 and parameters that are suitable for use by a waveform synthesizer 142. The waveform synthesizer 142 is the synthesis portion of a vocoder which requires thirteen parameters to be provided every ten milliseconds: one describing the fundamental frequency of the speech, one describing the frequency of the voiced/unvoiced bands, one describing the total energy of the ten millisecond frame and ten line spectral frequency parameters describing the frequency spectrum of the frame. The output of the waveform synthesizer 142 is synthetic speech (neural network synthesized speech 144).

Similarly, the visual rendering module 122 requires that at least six three-dimensional coordinates are provided every 33.33 milliseconds. These coordinates describe the corners of the lips, the position of the top lip, the position of the bottom lip, the position of the chin and a reference position located on the nose. Preferably, these six points are supplemented with points describing positions of the eyebrows, forehead and cheeks for a total of seventeen reference points.

In addition, it is preferable to generate the degree to which the teeth and the tongue are visible. It is also preferable to generate parameters describing the direction that the eyes appear

to be looking and the frequency and speed of blinking the eyes. Together these points describeparameters describing degree of lip opening, lip protrusion, jaw position, tongue position, eye positions, eyebrow positions and parameters describing head position. Preferably, these are three-dimensional coordinates which correspond to nodes on a wire-frame model. Alternatively, the three-dimensional coordinates could also describe the wires connecting the nodes of the wire-frame model.

The visual rendering module 122 displays an animated image of a virtual actor (visually rendered model 124). The visually rendered model 124 and the neural network synthesized speech 144 are synchronized by the combining module 130 which generates a visually rendered model with speech 132 in order to convince the user that the speech is being generated by the virtual actor.

In order for the neural network module 112 to be useful, it must first be trained. FIG. 2 is a schematic representation of a system for creating a representative parameter vector database that is used to train the neural network module 112.

The parameter database 218 is generated by having a human speaker 202 generate speech. An audio and motion recording system 204 is used to record the speech and corresponding motions of the human speaker 202. The recordings are typically generated by placing two video cameras at a ninety-degree angle with respect to each other with one video camera set up to record the side view of the human's head and the other video camera is set up to record the front view of the human's head.

The audio is typically recorded by an audio recording device such as a digital audio tape recorder through a microphone placed in front of the human speaker 202. Alternatively, the audio can also be recorded by the video cameras or motion capture equipment can be used in place of or in addition to the video equipment.

Expression input 220 can be given to the human speaker 202 which is typically in the form of instructions such as ※ Please speak the following sentence in an angry tone of voice※ . The expression input 220 is converted to expression parameters 222 by the linguistic labeler 214. The linguistic labeler 214 also generates time aligned linguistic representation of speech 216 which are linguistic labels for each speech segment. The speech segments are typically phoneme segments and the linguistic labels are composed of the following: phoneme identifier associated with each phoneme; articulatory features associated with each phoneme; locations of syllables, words and other syntactic and intonational boundaries; duration of time between phonemes, syllables, words and other syntactic and intonational boundaries; syllable strength information; descriptive information of a word type; and prosodic information. The prosodic information is additionally composed of at least one of the following: the locations of word endings and the degree of disjuncture between words, the locations of pitch accents and the form of those accents, the locations of boundaries marked in the pitch contour and the form of those boundaries, the time separating marked prosodic events from other marked prosodic events, and the number of prosodic events of some type in the time

period separating a prosodic event of another type and the frame for which the coder parameters are being generated.

As stated above, the expression input 220 that was given to the human speaker 202 is converted to expression parameters 222. This conversion changes the expression input 220 (which is typically of the form ※ Please speak the following text in a happy tone of voice※) into a text label that is time aligned with the speech waveform (which for this example would read ※ happy※). The linguistic labeler 214 receives audio information 212 and expression input 220 and converts the audio information 212 in linguistic representation of speech 216 and converts the expression input 220 into predetermined labels (expression parameters 222). This process can also be done manually by a person who is skilled in the art of labeling speech.

Spatial parameters 210 are extracted from the motion information 206 by a spatial parameter extractor 208. If the motion of the human speaker 202 is recorded using a video recorder, then the spatial parameter extractor 208 is typically a tracking algorithm that locates key points of a video recording of the human face. White dots can be affixed or marked on the face at the key locations on the human speaker 202 to help the tracking algorithm locate these key points. Typically dots are placed on the lips, chin, eyebrows, forehead and cheeks with a total of approximately eight to fifty dots. The spatial parameter extractor 208 then combines the front and side views from the video recording and generates three-dimensional coordinates of the key points.

If the motion information 206 is generated by a motion capture device, then the markers or transmitters must be placed in the key locations as described above. In this case, the spatial parameter extractor 208 will have to do the appropriate processing necessary to turn the motion capture device output into three-dimensional coordinates. The spatial parameter extractor 208 generates spatial parameters 210 which can be the three-dimensional coordinates of the key points or, alternatively, parameters that are used to drive the visual rendering module 122.

The spatial parameters 210, linguistic representation of speech 216 and the expression parameters 222 all contain time stamps which are used to synchronize these components with each other. The spatial parameters 210, linguistic representation of speech 216 and the expression parameters 222 are stored in the parameter database 218 for later use.

Turning to FIG. 3, once the parameter database 302 has been generated, the spatial parameters 308, corresponding linguistic representation of speech 306 and optionally the corresponding expression parameters 310 are retrieved by the parameter retrieval module 304. The parameter retrieval module 304 insures that the correct spatial parameters 308, linguistic representation of speech 306 and optionally corresponding expression parameters 310 are retrieved such that they are correlated. Typically, this is done by insuring that the time stamps on each of these parameters are the same, though the various parameters may have different sampling rates. This means that spatial parameters may have thirty parameter vectors corresponding to a second of motion (when

the motion is recorded at thirty frames per second) whereas the linguistic representation of speech may have one hundred parameter vectors corresponding to a second of speech (where the speech is recorded at one hundred frames per second). Due to the potentially differing sampling rates, the parameter retrieval module 304 may have to interpolate between parameters in order to provide synchronized spatial parameters 308, linguistic representation of speech 306 and expression parameters 310.

The spatial parameters 308 are converted to raw spatial parameters 322 by a target processor 312. The raw spatial parameters 322 are neural network target parameters and are typically scaled versions of parameters which are suitable for driving a visual rendering module 122. This means that any scaling or conversion between spatial parameters extracted from a human speaker 202 and parameters necessary to drive the visual rendering module 122 are performed by the target processor 312. After the spatial parameters 308 have been scaled and converted to parameters that are suitable for driving a visual rendering module 122, the target processor 312 typically must scale the parameters into raw spatial parameters 322 which can be used to train a neural network. This typically means normalizing the vectors by scaling each element in the processed spatial parameter vector into a range whose minimum is zero and whose maximum is one.

Similarly, the linguistic representation of speech 306 is converted by the pre-processor 314 into neural network linguistic parameters 316 which are parameters that are suitable for input to a neural network module 320. In the preferred embodiment, the

neural network linguistic parameters contain four components, called streams.

As shown in FIG. 4, the first stream is called the breaks input 425 and contains twenty-six parameters describing the following: the degree of disjuncture between the previous word, the current word and the following word;
the duration of the previous, current and following words; the distances to the previous, current and following word boundaries; the distance to the highest fundamental frequency of speech in the current syntactic phrase; and the value of the highest fundamental frequency of speech for the current syntactic phrase.

The second stream is called the prosodic input 426 and contains forty-seven parameters describing the following: distances and values of phrase accents of the previous, current and future phrases; tone of the previous and current phrases; distances and values of the two previous, the two current and the two future pitch accents; total number of high pitch accents since the beginning of the phrase; and total number of down-stepped pitch accents since the beginning of the phrase.

The third stream is called the phonemic context 427 and contains sixty-three parameters describing the following: the current phoneme name, the prominence of the current word, the stress of the current syllable, and the category of the current word which is preferably a part-of-speech tag.

The fourth stream is called the duration/distance input 428 and contains three hundred and sixty-eight parameters which describe the following: duration of the previous five phoneme

boundaries; distances to the previous five phoneme boundaries, sum of one divided by the frame number for all frames of each of the previous five phonemes, duration to the following five phoneme boundaries, distances to the following five phoneme boundaries, sum of one divided by the frame number for all frames of each of the following five phonemes, and a description of the syntactic boundaries.

The expression parameters 310 are converted to neural network expression parameters 318 by the pre-processor 314. The neural network expression parameters can be contained by a separate fifth stream or alternatively combined with the prosodic input 426 in the second stream. In the preferred embodiment, the processed expression parameters 318 contain six parameters to describe the degree of the six basic expressions. The processed expression parameters 318 are preferably added to the stream containing the prosodic input 426.

In detail, FIG. 4 displays the architecture of the neural network module 112 in accordance with the present invention. The neural network is composed of a layer of processing elements with a predetermined specified activation function and at least one of the following: another layer of processing elements with a predetermined specified activation function, a multiple layer of processing elements with predetermined specified activation functions, a rule-based module that generates output based on internal rules and input to the rule-based module, a statistical system that generates output based on input and an internal statistical function and a recurrent feedback mechanism. The

neural network is hand modularized according to speech domain expertise.

The neural network contains two phoneme-to-feature blocks 402 and 403 which use rules to convert the unique phoneme identifier contained in both the phonemic context 427 and the duration/distance input 428 to a set of acoustic features such as sonorant, obstruant and voiced. The neural network also contains a recurrent buffer 415 which is a module that contains a recurrent feedback mechanism. This mechanism stores the output parameters for a specified number of previously generated frames and feeds a non-linear sampling of the previous output parameters back to other modules which use the output of the recurrent feedback mechanism 415 as input.

The square blocks in FIG. 4 404-414 and a 416-420 are modules which contain a single layer of perceptrons. The neural network input layer is composed of several single layer perceptron modules 404-409 which have no connections between each other. All of the modules in the input layer feed into the first hidden layer 410. The output from the recurrent buffer 415 is processed by a layer of perceptron modules 416-418. The information from the output of the recurrent buffer, the recurrent buffer layer of perceptron modules 416-418 and the output of the first hidden layer 410 is fed into a second hidden layer 411, 412 which in turn feeds the output layers 413 and 414.

In the preferred embodiment, the neural network module 112 generates both speech and spatial parameters from the same input. The output contains two modules 419 and 420 to generate the speech

parameters and two modules 413 and 414 to generate the spatial parameters. The first speech output module 419 computes the voicing and energy parameters for the speech frame, whereas the second speech output module 420 generates the line spectral frequencies which describe the spectral contour of the speech frame. The first spatial parameter output module 413 generates a first set of reference points which are the frequency and degree of oscillations for spatial parameters, whereas the second spatial parameter output module 414 generates a second set of reference points which are the coordinates of the spatial parameters, which later get converted to three-dimensional coordinates that are used to drive the visual rendering module 122. The second set of reference points are different from the first set of reference points. The neural network output module 401 combines the output from each module in the output layer 413, 414, 419 and 420 and converts the information into parameters that are suitable for use by external devices.

FIG. 4 also shows tapped delay line input modules 421-424 which can be used to provide a non-linear context window of the input streams. In the preferred embodiment, a tapped delay line module 423 is used to provide phonemic context 427 of a total of thirty prior and future phonemic context vectors. These are not just the thirty adjacent phonemes that are presented but are the vectors from frames -40, -36, -32, -28, -24, -20, -16, -12, -8, -6, -4, -2, -1, 0, 1, 2, 3, 4, 5, 7, 9, 11, 15, 19, 23, 27, 31, 35, 39 and 43, where the negative frame numbers indicate previous frames and positive frame numbers indicate the future frames.

Each frame has a preferable duration of five milliseconds. Alternatively the breaks input 425, prosodic input 426 and duration/distance input 428 can be processed by tapped delay line modules 421, 422 and 424, respectively.

Since the number of neurons is necessary in defining a neural network, the following table shows the details about each module:

FIG. number	Module Type	Number of Inputs	Number of Outputs
401	rule	70	70
402	rule	2280	1680
403	rule	438	318
404	single layer perceptron, sigmoid activation	26	15
405	single layer perceptron, sigmoid activation	47	15
406	single layer perceptron, sigmoid activation	2280	15
407	single layer perceptron, sigmoid activation	1680	15
408	single layer perceptron, sigmoid activation	446	15

409	single layer perceptron, sigmoid activation	318	10
410	single layer perceptron, sigmoid activation	85	120
411	single layer perceptron, sigmoid activation	128	30
412	single layer perceptron, sigmoid activation	160	40
413	single layer perceptron, sigmoid activation	40	51
414	single layer perceptron, sigmoid activation	45	6
415	recurrent mechanism	70	700
416	single layer perceptron, sigmoid activation	700	5
417	single layer perceptron, sigmoid activation	700	10

418	single layer perceptron, sigmoid activation	700	20
419	single layer perceptron, sigmoid activation	40	3
420	single layer perceptron, sigmoid activation	45	10

The neural network is trained using the above procedure by using a back propagation of errors algorithm. A gradient descent technique or a Bayesian technique may alternatively be used to train the neural network.

Alternatively, the linguistic representation of speech can be the output of a speech recognizer 105. The speech recognizer would have speech 103 as input and would generate the linguistic parameters described above. The neural network would preferably be trained from more than one speaker and the pre-processor 108, neural network module 112 and post-processor 116 would all preferably receive speaker profile information which would be generated by the speech recognizer. The speaker profile is coupled to at least one of the text-to-linguistics module, the neural network based parameter system and the visual rendering module. Moreover, the speaker profile has at least one of the following: visual profiles and voice profiles (the visual

profiles are associated with the voice profiles determined by a user identifier).

There are two methods of using the speaker profile for the neural network based parameter system 138. The speaker profile would preferably be used to select predetermined weights for the neural network but alternatively the speaker profile could be used to change predetermined inputs which would effect the output even though the weights would not change. In this alternative method, the neural network would have to be trained using the predetermined inputs which would remain at constant values for the training data that represents one human. The values would then be altered to a new predetermined constant value for a new human that is used to generate data for the parameter database 218 and 302.

The pre-processor 108 and the post-processor 116 may also receive the speaker profile for the neural network based parameter system 138. The pre-processor 108 changes at least one of the predetermined neural network expression parameters 148 and the neural network linguistic parameters 110 based on the speaker profile. The post-processor 116 uses the speaker profile to modify the predetermined model parameters 120. In the preferred embodiment, the speaker profile is also used by the neural network module 112 to alter the parametric representation of speech 134 which ultimately changes the characteristics of the neural network synthesized speech 144. This is preferably done by changing the weights of the neural network according to the speaker profile.

The speaker profile for the visual rendering module 140 may contain information such as age, skin color, hair color, eye color

and dimensions of facial features. Whenever these general setup parameters are updated, the visual rendering module 122 will cause the visually rendered model 124 to change its appearance in a predetermined way.

The text-to-linguistics module 104 preferably receives a speaker profile for text-to-linguistics module 134 which would include at least one of language identifier, socio-linguistic information and physical information relevant to speech production. This speaker profile causes the text- to-linguistics module 104 to alter its output in a predetermined manner. For example, if the speaker profile identified that the speaker was from England, then the linguistic representation of speech 106 would contain British pronunciation of words in the text 102.

In the preferred embodiment, the text-to-linguistics module 104 also generates expression parameters 146 which are preferably derived from the text 102. The expression parameters 146 describe expressions such as sadness, anger, joy, fear, disgust or surprise. Alternatively, these expression parameters 146 might be extracted from text 102 that is written with some markup language. As an example, the text 102 might have the following sentence: ※ I am <emphasize> very late < \emphasize>.※ The text-to-linguistics module 104 would generate the expression parameter 146 EMPHASIZED which would be active during the duration of the phrase "very late※ . These expression parameters 146 effect the model parameters 120 used to generate the visually rendered model 124 and one of the parametric representation of speech 134 and the

alternatively synthesized speech 128 used to generate the visually rendered model with speech 132.

An alternative embodiment uses a non-neural network based linguistics-to-speech module 126 to convert the linguistic representation of speech 106 into alternatively synthesized speech 128. In this embodiment, the raw speech parameters 115, parametric representation of speech 134, waveform synthesizer 142 and neural network synthesized speech 144 would not be present. The speaker profile for linguistics-to-speech module 126 would provide information which would change the characteristics of the alternatively synthesized speech 128. As an example, the speaker profile might contain information such as the gender of the speaker which would effect the pitch and spectral parameters of the alternatively synthesized speech. The linguistics-to-speech module 126 may be combined with the text-to-linguistics module 104 in at least one of a speech synthesis system, text-to-speech system and a dialog system.

In the preferred embodiment the coding of the head position in terms of horizontal rotation, vertical rotation and tilt is coded as an amplitude and frequency of oscillation. The amplitude describes the limits or degree of movement and the frequency describes the speed of the movement. Tilt is defined as the rotation about the horizontal axis which passes through the front and back of the head and approximately intersects the bridge of the nose. The combination of the three rotation parameters defines the direction that the head is facing.

Other non-articulatory gestures such as blinking, smiling, shoulder shrugs, head nodding, winking and coughing are also generated by the neural network-based parameter system 118 in the preferred embodiment. These events are preferably coded by using the method of generating the amplitude and frequency of the events, as described above. The speaker profile for neural network-based parameter system 138 and the speaker profile for the visual rendering module 140 preferably contains parameters describing which of the predetermined events are desired to be displayed in the visually rendered model 124 and the degree at which they should be present. Note that the device of the present invention may be implemented in a text-to-speech system, a speech synthesis system or a dialog system.

A general purpose computer implementing the present invention is shown in FIG. 5. A computer monitor 510 having a screen 512 displaying a virtual actor 514, a central processing unit 516 having a microprocessor 518, a keyboard 520 and a data storage medium 522 (e.g., computer discs, floppy discs, etc.) are shown. The data storage medium 522 has stored thereon a set of instructions which, when loaded into the computer causes the computer to act as a neural network. When the computer is stimulated by linguistic parameters, the linguistic parameters are converted into raw spatial parameters. The storage medium 522 further has stored thereon a set of speaker profiles 134, 136, 138 and 140, including at least visual profile data and voice profile data. When the visual profile data and the voice profile data are loaded into the computer, they cause the computer to display the

virtual actor 514 having a visual appearance determined by the visual profile data and a voice determined by the voice profile data.

Alternatively, the neural network based parameter system 118 can be driven by an acoustic representation of speech, giving, as its output, spatial parameters which include model parameters (facial parameters and gestures of the face and body such as, eyebrow movement, head tilt, shoulder shrugs, hand waving, etc.). These spatial parameters can be in addition to spatial parameters of lip movement. The acoustic representation of speech would preferably come from a speech recognition device but could come from a speech coder. The neural network is trained using the acoustic representation of speech as an input and generates the raw spatial parameters as described above. Instead of containing information on a select number of speakers, the parameter database would contain information recorded from a large predetermined number of human speakers.

The present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive. The scope of the invention is, therefore, indicated by the appended claims rather than by the foregoing description. All changes which come within the meaning and range of equivalency of the claims are to be embraced within their scope.

We claim:

Claims

1. A data storage medium having stored thereon a set of speaker profiles, including at least visual profile data and voice profile data, which, when loaded into a computer causes the computer to display a virtual actor having a visual appearance determined by the visual profile data and a voice determined by the voice profile data.

2. A computer which, when loaded with the set of speaker profiles of claim 1, displays a virtual actor, wherein the appearance of the virtual actor is selectable dependent on selection of a speaker profile from the set of speaker profiles.

Application No: GB 0009707.1
Claims searched: 1

Examiner: Leslie Middleton
Date of search: 1 June 2000

Patents Act 1977
Search Report under Section 17

Databases searched:

UK Patent Office collections, including GB, EP, WO & US patent specifications, in:

UK Cl (Ed.R): H4T (TABX)

Int Cl (Ed.7): G06T 1/40[6], 15/70

Other: ONLINE: EPODOC, JAPIO, WPI / EPOQUE

Documents considered to be relevant:

Category	Identity of document and relevant passage	Relevant to claims
X	EP 0696018 A (Nippon T & T) See p2(28-51);p4(1-24); p6(22-57);p10(23)-p11(26);p13(34-48)	1

X	Document indicating lack of novelty or inventive step	A	Document indicating technological background and/or state of the art.
Y	Document indicating lack of inventive step if combined with one or more other documents of same category.	P	Document published on or after the declared priority date but before the filing date of this invention.
&	Member of the same patent family	E	Patent document published on or after, but with priority date earlier than, the filing date of this application.